View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by TELKOMNIKA (Telecommunication Computing Electronics and Control)

TELKOMNIKA, Vol.13, No.3, September 2015, pp. 1037~1046
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v13i3.1543

Burn Area Processing to Generate False Alarm Data for Hotspot Prediction Models

Imas Sukaesih Sitanggang*[1], Razali Yaakob[2], Norwati Mustapha[3], Ainuddin A. N.[4]
[1]Department of Computer Science, Faculty of Natural Science and Mathematics,
Bogor Agricultural University, Indonesia
[2,3]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia
[4]Institute of Tropical Forestry and Forest Products (INTROP), Universiti Putra Malaysia, Malaysia
*Corresponding author, e-mail: imas.sitanggang@ipb.ac.id[1], razaliy@upm.edu.my[2],
norwati@upm.edu.my[3], a_ainuddin@yahoo.com[4]

Abstract

Developing hotspot prediction models using decision tree algorithms require target classes to which objects in a dataset are classified. In modeling hotspots occurrence, target classes are the true class representing hotspots occurrence and the false class indicating non hotspots occurrence. This paper presents the results of satellite image processing in order to determine the radius of a hotspot such that random points are generated outside a hotspot buffer as false alarm data. Clustering and majority filtering were performed on the Landsat TM image to extract burn scars in the study area i.e. Rokan Hilir, Riau Province Indonesia. Calculation on burn areas and FIRMS MODIS fire/hotspots in 2006 results the radius of a hotspot 0.90737 km. Therefore, non-hotspots were randomly generated in areas that are located 0.90737 km away from a hotspot. Three decision tree algorithms i.e. ID3, C4.5 and extended spatial ID3 have been applied on a dataset containing 235 objects that have the true class and 326 objects that have the false class. The results are decision trees for modeling hotspots occurrence which have the accuracy of 49.02% for the ID3 decision tree, 65.24% for the C4.5 decision tree, and 71.66% for the extended spatial ID3 decision tree.

Keywords: *hotspot, satellite image processing, data mining, decision tree*

Copyright © 2015 Universitas Ahmad Dahlan. All rights reserved.

1. Introduction

Predictive models for hotspots occurrence are essential to develop so that damages caused by forest fires can be minimized. Nowadays, the large number of forest fire data has been triggered the development of data mining systems to analyze influencing factors for forest fires and their relations [1-5]. Data mining is a growing area in computer science that is widely used to extract interesting and valid information from large data. One of data mining techniques namely classification algorithms have been applied to model hotspots occurrence [6-8]. The task of classification aims to discover classification rules on a collection of objects which is represented in a relation (a dataset). The rules determine label classes of any object (Y) from the values of its attributes (X). Decision tree is one of famous methods in creating classification models. A decision tree is a model expressing classification rules which has three types of nodes i.e. a root node, internal, and leaf nodes. A root node or an internal node contains attribute test conditions to separate objects that have different characteristics. Leaf nodes hold the target classes (true class and false class) to which objects will be classified. In hotspots occurrence modeling, the classes are hotspots occurrence (True class) and non hotspots occurrence (False class). The attributes of objects may include some supporting factors for hotspots occurrence such as physical, socio-economic, as well as weather data. This study applied three decision tree algorithms i.e. ID3, C4.5 and extended spatial ID3 [9] on the forest fire dataset to develop models for classification and predicting hotspots occurrence.

Hotspots data are provided by several institutions such as NASA/University of Maryland and The ASEAN Specialised Meteorological Centre (ASMC). In addition to hotspots as true alarm data, a classification task in modeling hotspots occurrence requires non-hotspot points as false alarm data. This work aims to generate non-hotspot points near hotspots to prepare the target classes for modeling hotspots occurrence in Rokan Hilir District in Riau Province

Received February 10, 2015; Revised May 21, 2015; Accepted June 4, 2015

Indonesia. Burn area processing for the study area was performed to determine the radius of buffer for a hotspot and then outside of the buffer, we generated non-hotspot points. There are two main steps in burn area processing i.e. image classification and majority filtering. Image classification identifies classes on an image based on its spectral characteristics. In order to improve the accuracy of image classification, majority filtering is applied to remove very small areas resulted from the image classification.

Section 2 discusses materials and methods used in our study. The discussion includes the study area and the data utilized in this study. In addition, two methods in image processing are outlined in Section 2 namely classification and majority filtering. In Section 3, we present the results of burn area processing to generate false alarm data. The study is summarized in Section 4.

2. Materials and Methods

2.1. Study Area and Data

The study area is Rokan Hilir district in Riau Province in Indonesia (Figure 1). Rokan Hilir spans an area of 8,881.59 km^2 [10] or approximately 10% of Riau's total land area. The site is situated in the area between 100°16' - 101°21' East Longitude and 1°14' - 2°30' North Latitude [10]. Rokan Hilir is located in the western part of the north Sumatera, the southern part of Bengkalis district and Rokan Hulu district, the eastern of Dumai and the northern part of the north Sumatera and Malacca strait. According to [11], in 2002, Rokan Hilir had 454,000 hectares (ha) of p*eatlands* or about 11.2% of the whole peatlands in Riau Province.

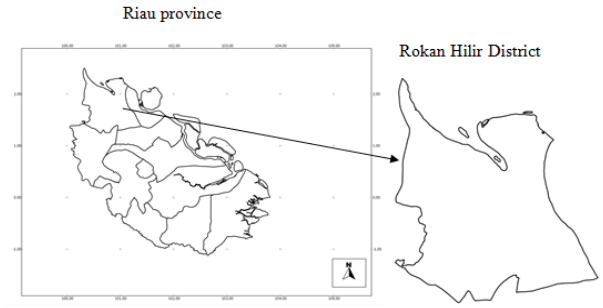


Figure 1. Study area

The data used in burn area processing are spread and coordinates of FIRMS MODIS fire/hotspots in 2006, as well as the Landsat TM image for extracting burn areas (Figure 2) (courtesy of the U.S. Geological Survey). The acquisition date of the image is 24 July 2006, the resolution of the image is 30×30 m^2 and the band combination used is 7, 4, 2. This combination is used in the fire management applications for post-fire analysis of burned and non burned forested areas. In Figure 2, the areas covered by white lines represent burn areas.

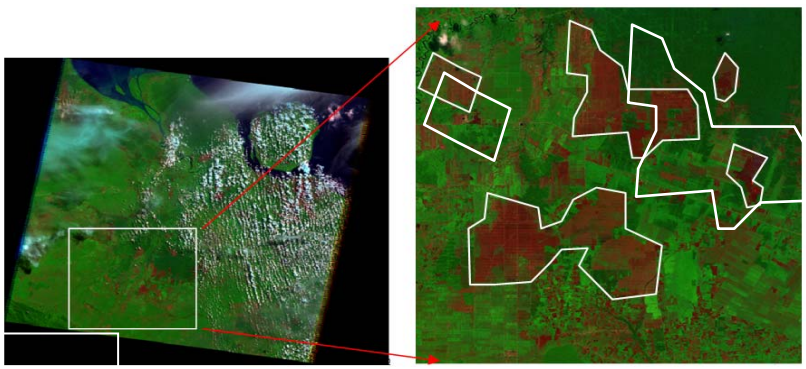


Figure 2. Landsat TM satellite imagery, Band combination 7, 4, 2

2.2. Tools for Data Processing

This work utilized ILWIS for processing the satellite image and Quantum GIS for spatial data processing and data visualization. The Integrated Land and Water Information System (ILWIS) is open source software for remote sensing and geographical information systems developed by the Faculty of Geo-Information Science and Earth Observation, of the University of Twente (http://www.ilwis.org/). Quantum GIS (QGIS) is a free and open source Geographic Information System. Several main features provided by QGIS include visualization, managing, editing, analysing spatial data, and composing printable maps (http://www.qgis.org).

2.3. Digital Image Processing

Digital image processing refers to a process that is conducted to improve an image. The purpose of this process is to assist the extraction of information about objects in a satellite image. Images in digital image processing are data acquired by remote sensors on satellite, aerial, or ground platforms. The images are available in the digital format with specific spatial, radiometric, and spectral characteristics.

A digital image is represented by a matrix in which each element in the matrix is denoted as a pixel (picture element). A pixel is associated with a Digital Number (DN), as well as rows and columns which determine the coordinate of the image. Reference [12] states that Digital Numbers (DNs) represent a discrete measure of the radiance (L) detected by the sensors and measured in Watts per square metre per steradian ($W \cdot m^{-2} \cdot sr^{-1}$). Actual physical measures of the radiation are continuously acquired and then the analogical/digital converters will alter these measures into discrete level [12]. In addition to DN values and the coordinate of the image, the spectral resolution is another essential characteristic of a satellite image. According to [12], the spectral resolution is the wavelength interval (λ) to which the radiance represented by its Digital Number refers. Several images can be available for the same scene to compose a multispectral image. Each image refers the radiance recorded in definite spectral ranges [12].

2.4. Image Classification

Image classification is a process to recognize classes on an image based on its spectral characteristics [12]. Classification tasks can be divided into two groups: unsupervised and supervised. In unsupervised classification, pixels in a dataset are clustered based on statistics only and the concept of distance (for example, Euclidean), without any user-defined training classes. This approach does not require external information for assigning the pixels to the different classes. K-Means clustering is the commonly used algorithm in unsupervised classification. In supervised classification, a priori knowledge about the classes for a sufficient number of pixels (training sets) is needed [12]. The training sets are prepared by an analyst based on his/her personal experience, previous knowledge about thematic maps, and in-field survey. Pixels in the supervised classification method are divided into two sets namely the training set and the test set. The training set is used to determine a classification model. The model is then utilized to classify objects in the test set. The successful supervised classification depends on the definition of classes to which the pixels should be assigned. Some techniques applied in the supervised classification include Neural Network and Support Vector Machines.

2.5. Majority Filtering

Majority filtering is a post-classification method to improve the accuracy of image classification. This method can reduce the “salt-and-paper” resulted from per-pixel classifiers. According to [13], the majority filter is determined by identifying a neighborhood structure and a threshold value. This method applies a moving window in which the majority class of pixels within the window is assigned to the central pixel [14]. The majority class of pixels is the most frequently occurring value of a pixel and its neighbors in the window. A standard majority filter which works in a 3×3 environment which considers 9 pixels in the input map (ILWIS (3.5) help 2008). The predominant value, i.e. mostly frequently occurring value, or class name is assigned to the center pixel in the output map. For example, 9 pixel values encountered in the input map is shown in Table 1 [15].

Table 1. Nine pixel values in the input map

9	3	9
11	5	7
7	7	13

The predominant value is 7. Therefore, the value for the output pixel is 7. The value or class name that is encountered first will be assigned to the center pixel as output if there is no predominant value can be found in the 9 pixel values.

2.6. Decision Tree Algorithms

Decision tree is one of widely used classification methods in data mining. A decision tree algorithm generates a tree model to classify objects to their classes based on the characteristics of the objects. A decision tree has three types of nodes: 1) a root node, 2) internal nodes, and 3) leaves or terminal nodes. The root node and internal nodes hold attribute test conditions to partition records that have different characteristics. Leaves nodes (terminals) store class labels of objects. Traversing a decision tree from the root node to the leaves nodes results a set of classification rules. The rules are utilized to describe characteristics of objects and to predict unknown class labels of objects.

The ID3 decision tree algorithm was developed by J. Ross Quinlan during the late 1970s and early 1980s. The algorithm has the principle, where it builds the tree in greedy manner starting from the root, and selecting most informative features at each step [16]. In order to select the best feature for splitting the set of objects, the algorithm calculates information gain. A feature with the highest information gain is selected as a splitting feature.

The C4.5 decision tree algorithm is a successor of ID3. The C4.5 algorithm uses also Information Gain to select optimal splitting attributes. This algorithm uses a different method called rule post-pruning. There are three main tasks in C4.5: 1) generate the tree using the ID3 algorithm, 2) convert the tree to a set of if-then rules, and 3) prune each rule by removing preconditions if the accuracy of the rule increases without it [16].

Both ID3 and C4.5 use information gain as a measure for attribute selection. The formula of information gain is calculated as follows. Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$ [17]. The entropy is a measure of expected information for classifying a tuple in D. The formula of entropy is as follows [17]:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

Where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$. The formula to calculate information needed after using A to split D into v partitions to classify D [17]:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j) \quad (2)$$

Information gain is defined as the difference between the original information requirement (i.e. based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A) [17].

$$Gain(A) = Info(D) - Info_A(D) \quad (3)$$

The extended ID3 algorithm is an improvement of the ID3 algorithm such that the algorithm can be directly applied on a spatial dataset containing a set of layers [9]. The algorithm uses spatial information gain to select the best layer for splitting the spatial dataset. The formula of spatial information gain is defined as follows [9]. Let a target attribute C in a target layer S has *I* distinct classes (i.e. c_1, c_2, …, c_I), entropy for S represents the expected information needed to determine the class of tuples in the dataset and defined as:

$$H(S) = -\sum_{i=1}^{I} \frac{SpatMes\ (S_{c_i})}{SpatMes\ (S)} \log_2 \frac{SpatMes\ (S_{c_i})}{SpatMes\ (S)} \quad (4)$$

SpatMes(S) represents spatial measure of layer S that can be area of intersection polygons or distance between two spatial features.

Let an explanatory attribute V in an explanatory (non-target) layer L has *q* distinct values (i.e. v_1, v_2, …, v_q). We partition the objects in target layer S according to the layer L then we have a set of layers L(v_i, S) for each possible value v_i in L. In our work, we assume that the layer L covers all areas in the layer S. The expected entropy value for splitting is given by:

$$H(S \mid L) = \sum_{j=1}^{q} \frac{SpatMes\ (L(v_j, S))}{SpatMes\ (S)} H(L(v_j, S)) \quad (5)$$

The spatial information gain for layer L is given by:

$$Gain(L) = H(S) - H(S|L) \quad (6)$$

Gain(L) denotes how much information would be gained by branching on the layer L. The layer L with the highest information gain, (Gain(L)), is chosen as the splitting layer at a node N in a spatial decision tree.

3. Results and Discussion

3.1. Clustering and Majority Filtering

The main purpose of burn area processing is to define the radius of a buffer for a hotspot such that random points as non-hotspots will be generated outside the buffer. There are two main steps in image processing: clustering, or unsupervised classification, to group pixels and majority filtering to remove very small areas. These two tasks were conducted using the tool Ilwis 3.7. To perform clustering and majority filtering, we determined the map subset for each band (band 7, 4, 2). The coordinates used to create a subset of the map are (631478.23, 166290.54) and (747008.03, 87449.25). Clustering was applied on the subset of image with the number of cluster is 15. Figure 3 shows the result of clustering on the subset of image.

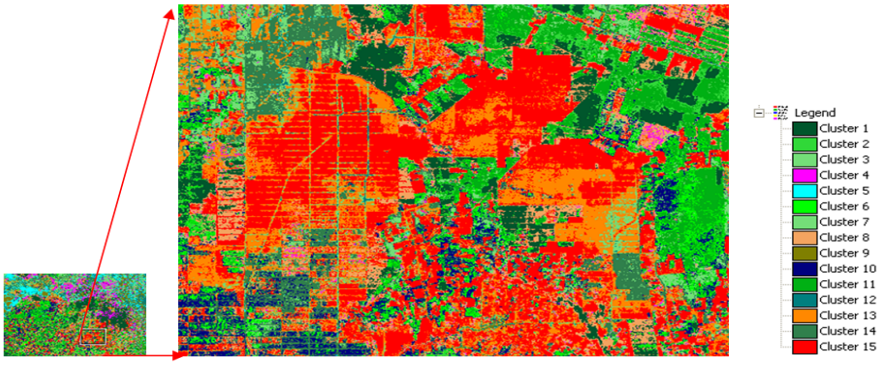
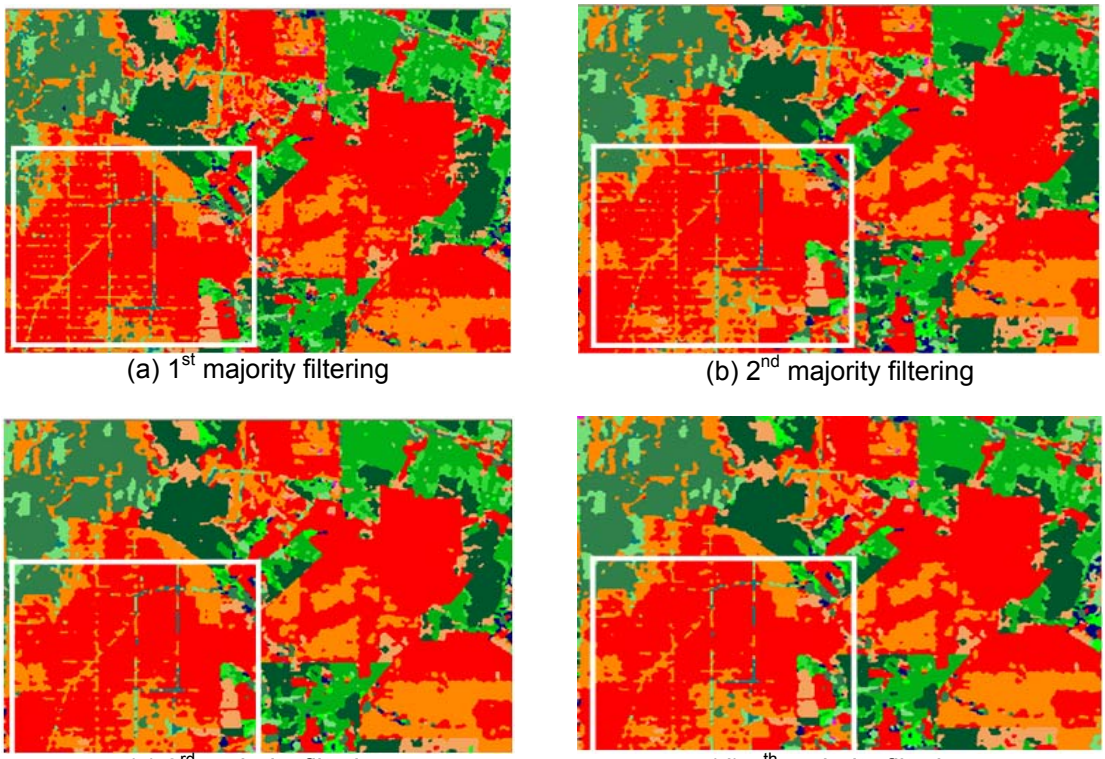


Figure 3. Clustering on the subset of image, number of cluster is 15

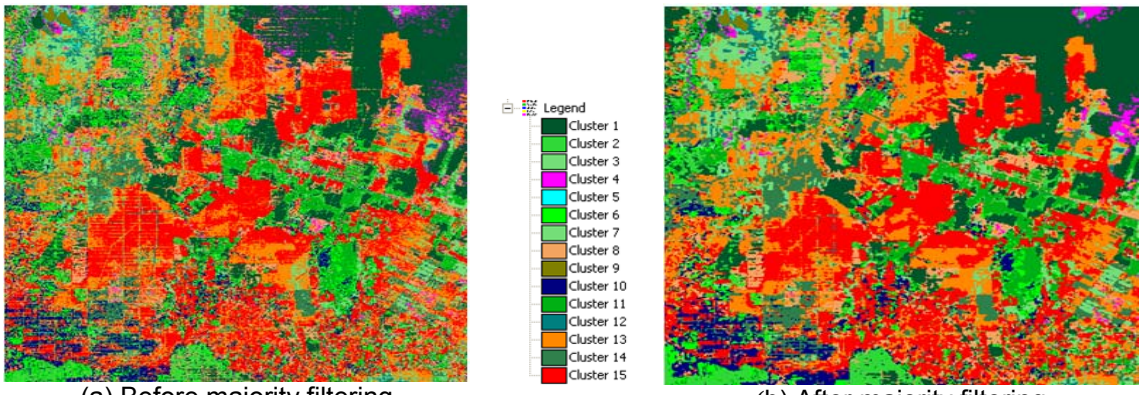
Furthermore, majority filter to remove very small areas was applied four times in the clustered image. The results are provided in Figure 4. The images resulted from the 1st and the 2nd majority filtering contain small areas as shown in the rectangular region. The small areas were reduced after we applied the 3rd and the 4th majority filtering. The images before and after applying majority filter are given in Figure 5. The use of majority filtering four times results the smoother image compared to those before applying majority filtering as shown in Figure 5.



(a) 1st majority filtering (b) 2nd majority filtering

(c) 3rd majority filtering (d) 4th majority filtering

Figure 4. Applying majority filtering on the image



(a) Before majority filtering (b) After majority filtering

Figure 5. Image before (a) and after (b) applying majority filtering

False alarms were generated outside buffers of hotspots as true alarm data using the tool Quantum GIS 1.7.2. The buffer operation that is available in Quantum GIS 1.7.2 is applied to point features (vector format). Therefore, the image in the raster format (tiff file) resulted from majority filtering was converted to the vector format (polygon). Figure 6 shows polygons only for bared lands (cluster 8), burn areas (cluster 13), and new burn areas (cluster 15).

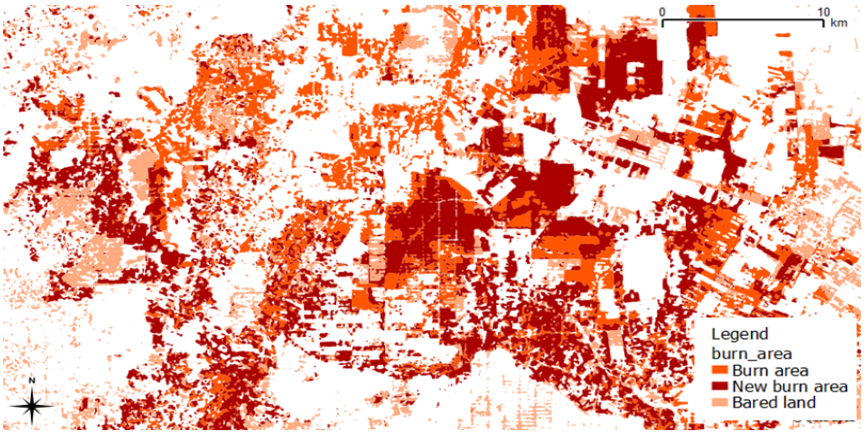


Figure 6. Burn areas and bared lands

In order to generate non hotspot points, this work involved only new burn areas (cluster 15). These burn scars were overlaid with hotspots that occurred in two weeks before the acquisition date for image (24 July 2006) (Figure 7). There were 298 hotspots in non-peatlands and one hotspot in peatlands found in the period 10 – 24 July 2006 in which 243 hotspots occurred in the burn scars.

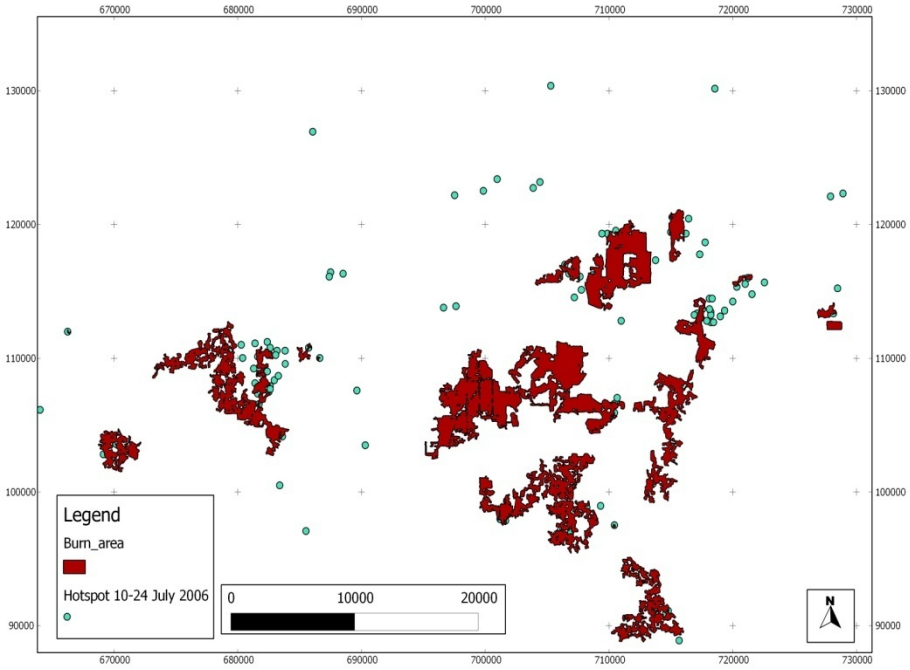


Figure 7. New burn areas and hotspots for the period 10 – 24 July 2006

To avoid single pixels labeling from the image, we consider only the burn scars with the area at least 1 ha that is equivalent to around 3×3 Landsat TM pixels. Therefore, burn scars with the area less than 1 ha were removed. This approach is also adopted in the work of [18]. Table 2 provides the summary of hotspots for the period 10-24 July 2006 in new burn areas at least 1 ha where the count of hotspots (associated with burn scars) is 243.

Table 2. Density of hotspots and area for one hotspot

	Area in km^2	Density (number of hotspot per km^2)	Area for one hotspot (Area in km^2/count of hotspot), in km^2
Max	44.75715	51.54616	11.18929
Average	9.78013	4.14282	**2.58657**
Min	0.01940	0.08937	0.01940
Sum	557.46733		

For simplicity, it is assumed that the area for a hotspot is a circle because a buffer of a hotspot is represented in a circle. The radius of the circle is given by $\sqrt{\text{area for one hotspot in km}^2/\pi}$ where π = 3.14159. As shown in Table 2, the area for one hotspot in average is 2.58657 km^2, therefore the radius of the circle is 0.90737 km. This value is considered as the radius of a buffer for a hotspot. Outside the buffers, random points are generated as false alarm data.

3.2. Generating Target Objects for Hotspot Prediction Models

As many 517 hotspots were found in Rokan Hilir in 2008. These hotspots were acquired by the MODIS satellite sensor. Buffers with the radius of 0.90737 km were created for each hotspot using Quantum GIS 1.7.2. Furthermore, as many 513 non hotspot points were randomly generated outside buffers. Therefore a non hotspot point is located at least 0.907374 km away from a hotspot (Figure 8). We consider these points as false alarm data which are combined to obtain target objects for the classification task.

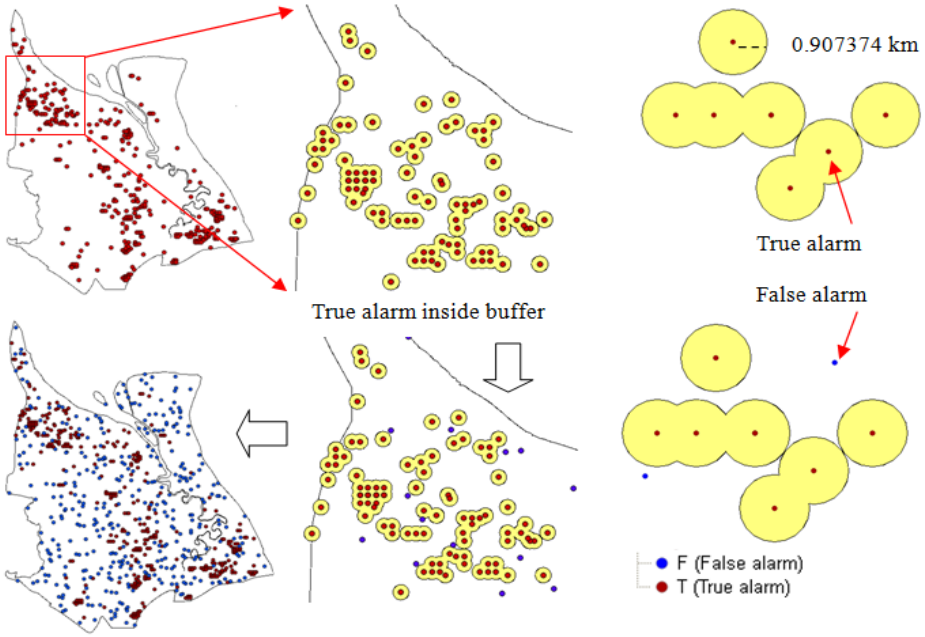


Figure 8. True and false alarm data as target objects

3.3. Predictive Models for Hotspots Occurrence

The decision tree algorithms namely ID3 and C4.5 have been applied on the forest fire dataset. Further discussion regarding these algorithms can be found in [19] and [16]. These algorithms are available in the data mining toolkit Weka 3.6.6. In addition, we created a model for predicting hotspots occurrence using our proposed method namely the extended spatial ID3 algorithm [9]. The algorithm is an improvement of the existing spatial ID3 algorithm introduced by [20]. Instead of running on the non-spatial dataset, our proposed algorithm works on the spatial dataset which contains several explanatory layers and one target layer. In a spatial database, layers stores spatial objects that can be represented either in point, line, or polygon.

Explanatory layers include layers of supporting factors for forest fires whereas the target layer consists of hotspots as true alarm data and non-hotspot points as false alarm data.

The spatial dataset for modeling hotspots occurrence has 1030 objects, one target layer and ten explanatory layers (distance to nearest city center (dist_city), distance to nearest river (dist_river), distance to nearest road (dist_road), income source, land cover, peatland type, peatland depth, precipitation in mm/day, screen temperature in K, 10m wind speed in m/s. In order to apply the ID3 and C4.5 algorithm, we conducted several steps to prepare a dataset from a spatial dataset on forest fires. These steps are as follows:

1. Calculating distance from target objects to nearest city center, river, and road
2. Relating layers that contain explanatory attributes and the target layer that consists of target classes.
3. Integrating all layers in by matching identifiers of objects to create a dataset for the classification task.
4. Remove duplicate objects in the dataset

Applying these steps on the spatial dataset on forest fires results 561 objects (235 true classes and 326 false classes). The experimental results show that the accuracy of ID3 decision tree is 49.02% and the accuracy of C4.5 decision tree is 65.24%. Furthermore, in term number of rules generated from the trees, the C4.5 algorithm outperforms the ID3 algorithm. The ID3 algorithm has 270 leaves with peatland type as the first test attribute whereas the C4.5 algorithm produces only 35 rules and the first test attribute of the tree is peatland type. The C4.5 decision tree has several test attributes to classify the objects to the target classes, i.e. peatland type, distance to nearest road, distance to nearest city center, screen temperature, distance to nearest river, and income source. For comparison, our proposed algorithm (Sitanggang et al. 2011) generated a spatial decision tree with 134 leaves and the first test layer of the tree is income source. The spatial decision tree has higher accuracy than the ID3 and C4.5 decision trees i.e. 71.12%. After pruning, the spatial decision tree becomes smaller with 122 leaves and its accuracy is 71.66%.

4. Summary

This work processed burn areas in the study area to generate non hotspot points as false alarm data in modeling hotspot occurrence models. Processing on the Landsat TM image and FIRMS MODIS fire/hotspots in 2006 shows that the area for one hotspot in average is 2.586562389 km^2. Therefore with the assumption that the area for a hotspot is a circle, the radius of a buffer is 0.907374 km. Experiments on the forest fires dataset result three decision tree models for hotspots occurrence prediction. The dataset contains influencing factors for forest fires, hotspots as true alarm data and non-hotspots as false alarm data. The three models are the ID3 decision tree with the accuracy of 49.02%, the C4.5 decision tree with the accuracy of 65.24% and the spatial decision tree with the accuracy of 71.66%.

Acknowledgements

The authors would like to thank Indonesia Directorate General of Higher Education (IDGHE), Ministry of National Education, Indonesia for supporting PhD Scholarship (Contract No. 1724.2/D4.4/2008) and Southeast Asian Regional Center for Graduate Study and Research in Agriculture (SEARCA) for partially supporting the research.

References

[1] Tay SC, Hsu W, Lim KH, Yap LC. *Spatial data mining: clustering of hot spots and pattern recognition*. Paper presented at the IEEE International Geoscience and Remote Sensing Symposium (IGARSS'03). Toulouse. 2003.

[2] Yu L, Bian F. *An Incremental Data Mining Method for Spatial Association Rule in GIS Based Fireproof System*. Paper presented at the International Conference on Wireless Communications, Networking and Mobile Computing, (WiCom 2007). Shanghai, China. 2007.

[3] Prasad KSN, Ramakrishna S. An Autonomous Forest Fire Detection System based on Spatial Data Mining and Fuzzy Logic. *International Journal of Computer Science and Network Security.* 2008; 8(12): 49-55.

[4] Hu L, Zhou G, Qiu Y. *Application of Apriori Algorithm to the Data Mining of the Wildfire*. Paper presented at the 6th international conference on Fuzzy systems and knowledge discovery. Tianjin, China. 2009.
[5] Angayarkkani K, Radhakrishnan. Efficient Forest Fire Detection System: A Spatial Data Mining and Image Processing Based Approach. *International Journal of Computer Science and Network Security.* 2009; 9(3): 100- 107.
[6] Stojanova D, Panov P, Kobler A, Džeroski S, Taškova K. *Learning to Predict Forest Fires with Different Data Mining Techniques*. Paper presented at the conference on Data Mining and Data Warehouses. Ljubljana, Slovenia. 2009.
[7] Sitanggang IS, Ismail MH. Classification model for hotspot occurrences using a decision tree method. *Geomatics, Natural Hazards and Risk*. 2011; 2(2): 111-121.
[8] Sitanggang IS, Yaakob R, Mustapha N, Nuruddin AN. Application of classification algorithms in data mining for hotspots occurrence prediction in Riau Province Indonesia. *Journal of Theoretical and Applied Information Technology*. 2012; 43(2): 214-221.
[9] Sitanggang IS, Yaakob R, Mustapha N, Nuruddin AN. *An extended ID3 decision tree algorithm for spatial data*. Paper presented at the IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM). Fuzhou, China. 2011.
[10] Rokan Hilir District. Overview of District. last modified 2009. Accessed May 30, 2012. http://www.rohilkab.go.id/?tampil=linkandact=profilandid=4.
[11] Wahyunto, Ritung S, Suparto, Subagjo H. *Peatland distribution and its carbon content in Sumatera and Kalimantan*. Project of Climate Change, Forests and Peatlands in Indonesia. Bogor: Wetlands International – Indonesia Programme and Wildlife Habitat. Canada. 2005.
[12] Gomarasca M. *Basics of Geomatics*. New York: Springer. 2009.
[13] KIM KE. 1996. Adaptive Majority Filtering for Contextual Classification of Remote Sensing Data. *International Journal of Remote Sensing.* 1996; (17): 1083-1087.
[14] Canty Morton J. Image Analysis, Classification, and Change Detection in Remote Sensing: with Algorithms for ENVI/IDL. Boca Raton: CRC Press. 2010: 237-238.
[15] ILWIS 3.4 Open. ILWIS (3.4) Help. last modified 2008. Accessed July 13, 2011, http://spatial-analyst.net/ILWIS/help.html.
[16] Marsland S. Machine Learning: An Algorithmic Perspective. Chapman & Hall/CRC machine learning & pattern recognition series. Boca Raton: CRC Press. 2009.
[17] Han J, Kamber M. Data Mining: Concepts and Techniques. Second edition. The Morgan Kaufmann series in data management systems. San Francisco: Morgan Kaufmann. 2006.
[18] Tansey K, Beston J, Hoscilo A, Page SE, Paredes Hernández CU. Relationship between Modis Fire Hot Spot Count and Burned Area in a Degraded Tropical Peat Swamp Forest in Central Kalimantan, Indonesia. *Journal of Geophysical Research*. 2008; 113 (D23): 1-8.
[19] Quinlan JR. Induction of decision trees. *Machine Learning.* 1986; 1(1): 81-106.
[20] Rinzivillo S, Franco T. *Classification in Geographical Information Systems*. Paper presented at the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases. Pisa, Italy. 2004.